Rapport - Prédiction de Structure Secondaire

KEDDIS Adam

March 11, 2024

Contents

1 Introduction

Le calcul de la structure d'une protéine est un problème classique en bio-informatique. Aujourd'hui considéré comme un dogme : *La structure permet la fonction*. La compréhension du repliement permettrait d'identifier la fonction d'une protéine à partir de sa séquence en acide aminé (AA à partir d'ici). Avec un regard porté sur l'avenir, même générer des séquences répondant aux structures désirées et donc à la fonction désirée. Cela permettrait par exemple la conception d'outils moléculaires très fins ou encore de médicaments.

Le calcul de la structure d'une protéine est donc un problème fondamental, qui une fois résolu, ouvrirait de nombreuses portes dans tous types de domaines.

Ainsi, dans ce projet, nous nous intéressons en particulier à la structure secondaire et plus précisément, aux hélices et aux feuillets.

2 Projet

Le projet consiste en l'analyse de séquences protéiques afin de pouvoir prédire la structure secondaire. Ici, on ne s'intéresse qu'aux hélices et aux feuillets. Dans un premier temps, on ne réalisera que l'analyse de différentes variables, avant d'utiliser des algorithmes de fouille de données et d'apprentissage.

2.1 Données

Pour les données, il y a 100 protéines d'*Escherichia coli* dont la séquence et les structures secondaires ont été récupérées sur la Protein Data Bank (PDB).

Le set de test contient également 30 protéines d'*Escherichia coli*, également récupérées sur la PDB.

2.2 Les variables

On s'intéressera aux variables des acides aminés. Voici le tableau recensant les variables utilisées :

	Polarité	Charge à pH 7	Volume	Tendance Hélice
A	0	0	11.5	1.489
R	52	+1	14.28	1.224
N	3.38	0	12.82	0.772
D	49.7	-1	11.68	0.924
C	1.48	0	13.46	0.966
Q	3.53	0	14.45	1.164
E	49.9	-1	13.57	1.504
G	0	0	3.4	0.510
H	51.6	0	13.69	1.003
I	0.13	0	21.4	1.003
L	0.13	0	21.4	1.236
K	49.5	+1	15.71	1.172
M	1.43	0	16.25	1.363
F	0.35	0	19.8	1.195
P	1.58	0	17.43	0.492
S	1.67	0	9.47	0.739
T	1.66	0	15.77	0.785
W	2.1	0	21.67	1.090
Y	1.61	0	18.03	0.787
V	0.13	0	21.57	0.990
Source	Zimmerman	Convention	Zimmerman	Deleage

	Tendance Feuillet	Flexibilité	Hydropathie	Pi
A	0.709	0.36	1.8	6
R	0.92	0.53	-4.5	10.76
N	0.604	0.46	-3.5	5.41
D	0.541	0.51	-3.5	2.77
C	1.191	0.35	2.5	5.05
Q	0.84	0.49	-3.5	5.65
E	0.567	0.5	-3.5	3.22
G	0.657	0.54	-0.4	5.97
H	0.863	0.32	-3.2	7.59
I	1.799	0.46	4.5	6.02
L	1.261	0.37	3.8	5.98
K	0.721	0.47	-3.9	9.74
M	1.21	0.3	1.9	5.74
F	1.393	0.31	2.8	5.48
P	0.354	0.51	-1.6	6.3
S	0.928	0.51	-0.8	5.68
T	1.221	0.44	-0.7	5.66
W	1.306	0.31	-0.9	5.89
Y	1.266	0.42	-1.3	5.66
V	1.965	0.39	4.2	5.9
Source	Deleage	Bhaskaran	Kyte	Zimmerman

On définit ainsi une séquence (SEQ) comme étant un mot w, définit sur Σ = {A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S ,T, W, Y, V}
avec w $\in \Sigma^*$.

Ainsi pour une variable (var), elle est calculée comme suit:

$$\text{SEQ[var]} = \frac{\sum_{AA \in \text{SEQ}} AA[var]}{|SEQ|} \tag{1}$$

2.3 L'analyse

Grâce aux structures sur la PDB, la séquence est récupérée et est étiquetée en Hélice ou en Feuillet, ce qui nous servira pour le K-NN. Ensuite, l'étude des variables est entreprise pour chaque séquence.

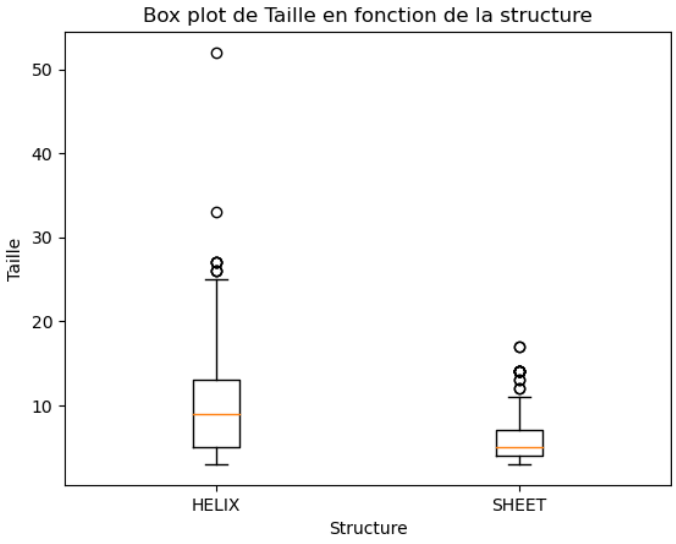
3 Résultats

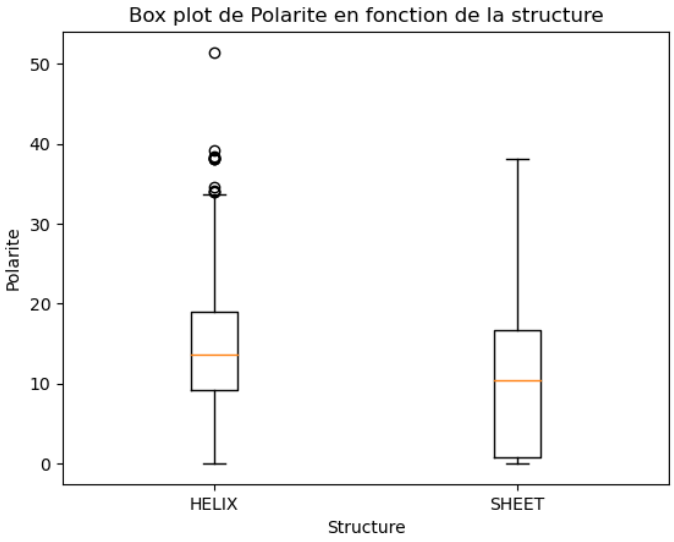
Les résultats de chaque sous section est présentée.

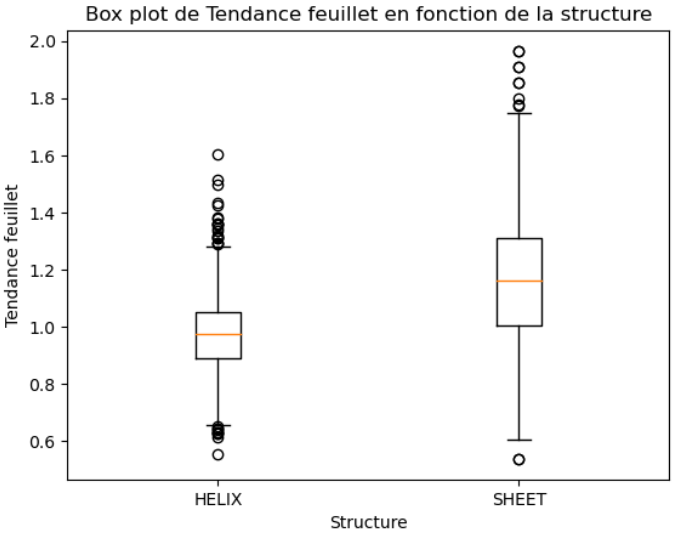
3.1 Les variables

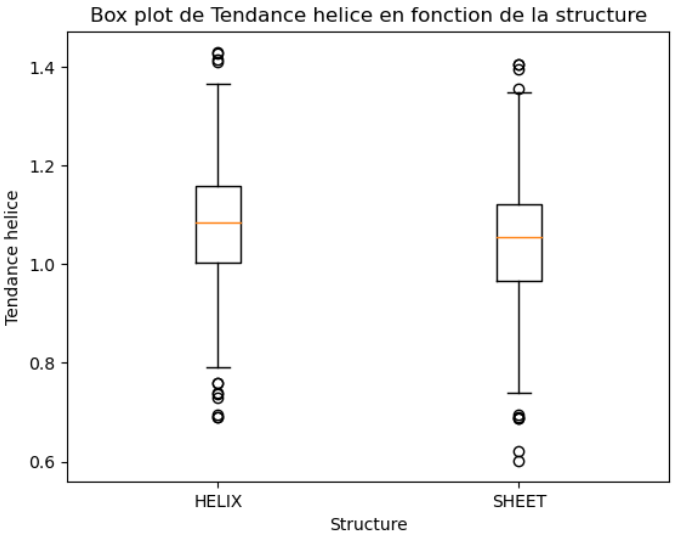
Seulement les variables présentant un Δvar important entre Hélice et Feuillet sont présentées.

Boxplot

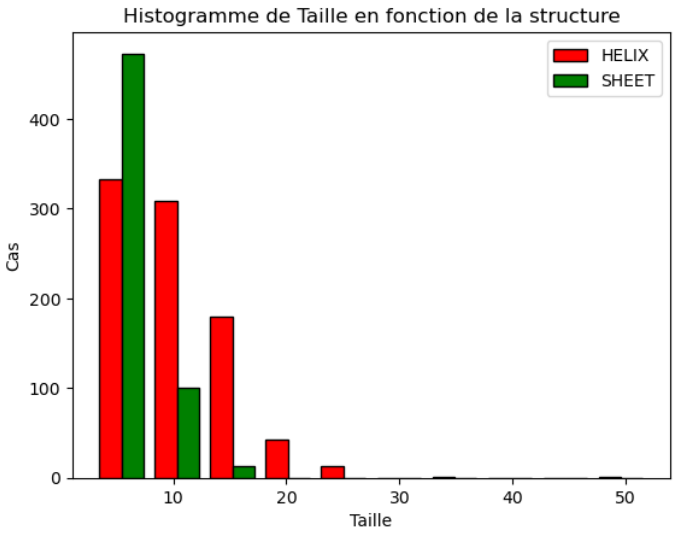


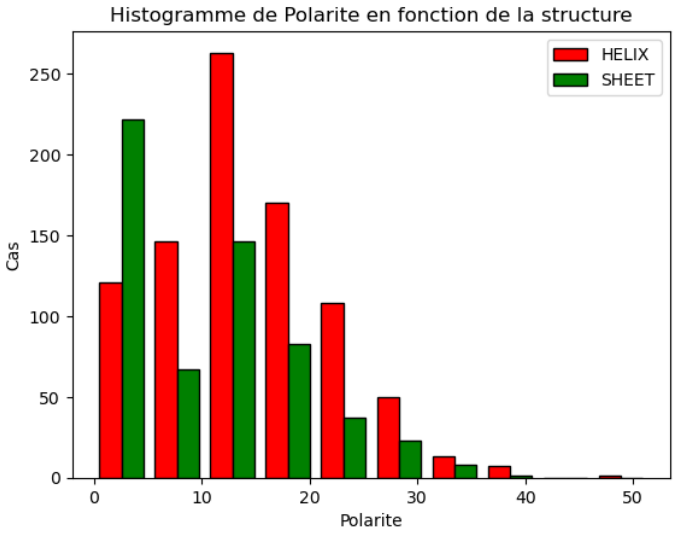

Box plot de Polarite en fonction de la structure
Polarite
50
40
30
20
10
0
HELIX
SHEET
Structure

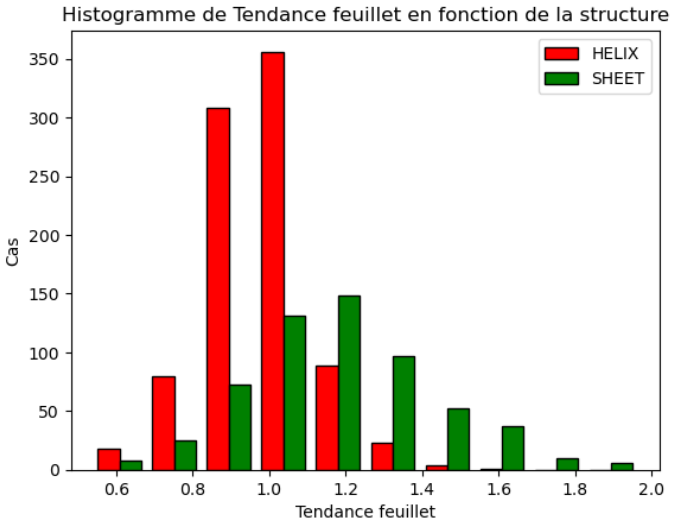

Box plot de Tendance feuillet en fonction de la structure
Tendance feuillet
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
HELIX
SHEET
Structure

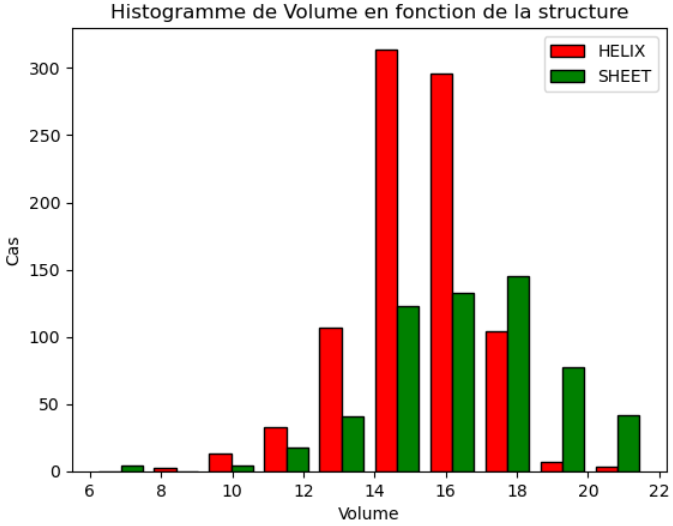

Box plot de Tendance helice en fonction de la structure
Tendance helice
1.4
1.2
1.0
0.8
0.6
HELIX
SHEET
Structure

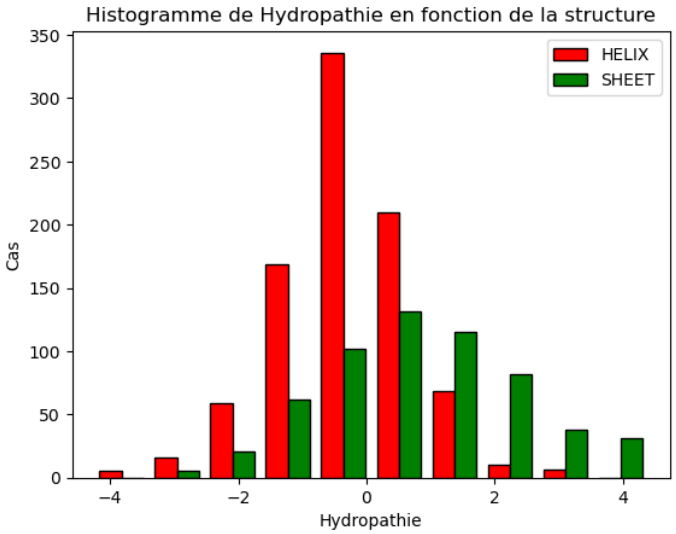
Histogramme






Histogramme de Tendance feuillet en fonction de la structure
HELIX
SHEET
Cas
Tendance feuillet
0
50
100
150
200
250
300
350
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0


Histogramme de Volume en fonction de la structure
HELIX
SHEET
Cas
Volume
0
50
100
150
200
250
300
6
8
10
12
14
16
18
20
22


Histogramme de Hydropathie en fonction de la structure
HELIX
SHEET
Cas
Hydropathie
350
300
250
200
150
100
50
0
−4
−2
0
2
4

3.2 Les corrélations

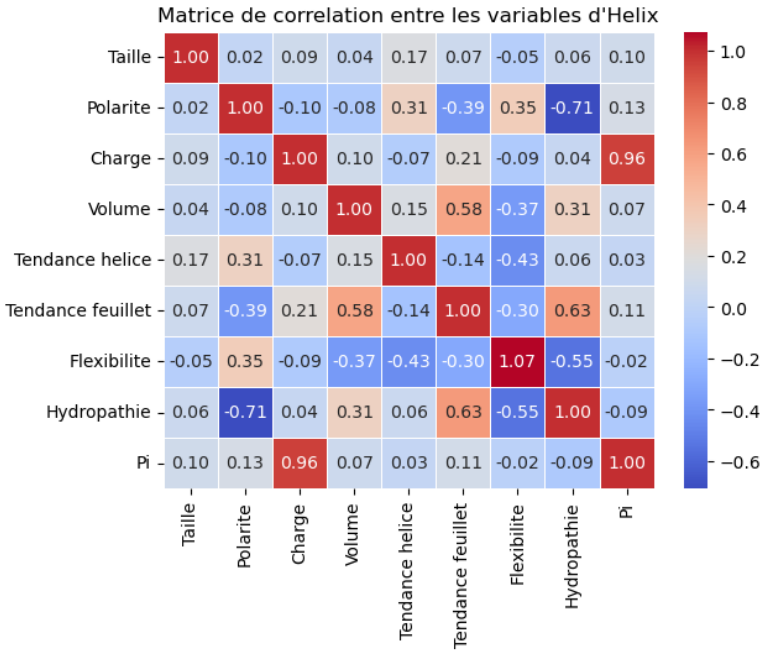
Les corrélations sont calculées à l'aide du coefficient de corrélation de Pearson:

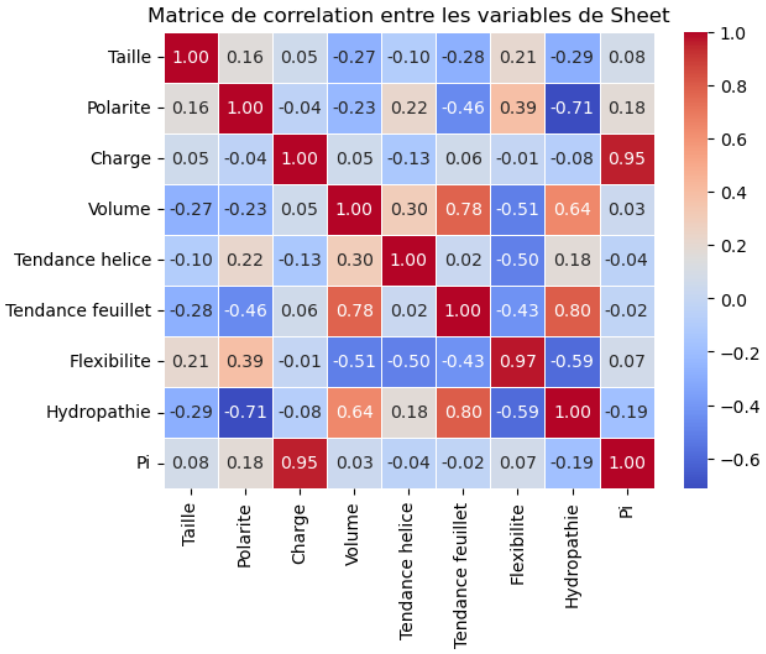
$$\text{Corr(var1, var2, str)} = \frac{\sum_{SEQ \in \text{str}} (SEQ[var1] - Moy(var1, str))(SEQ[var2] - Moy(var2, str))}{(|str| - 1) * \sqrt{Var(var1, str)} * \sqrt{Var(var2, str)}} \tag{2}$$

$$\text{Moy(var, str)} = \frac{\sum_{SEQ \in \text{str}} SEQ[var]}{|str|} \tag{3}$$

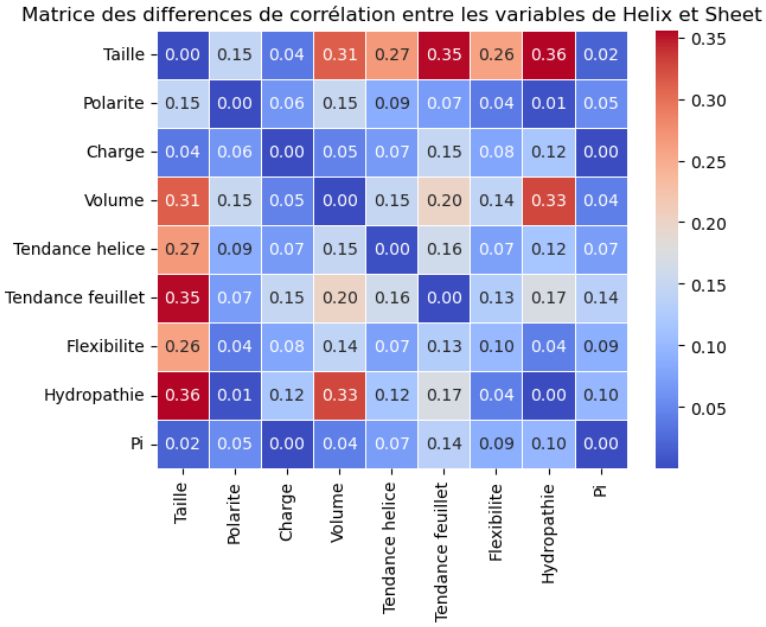
$$\text{Var(var, str)} = \frac{\sum_{SEQ \in \text{str}} (SEQ[var] - Moy(var, str))^2}{|str| - 1} \tag{4}$$

Ainsi, on a:





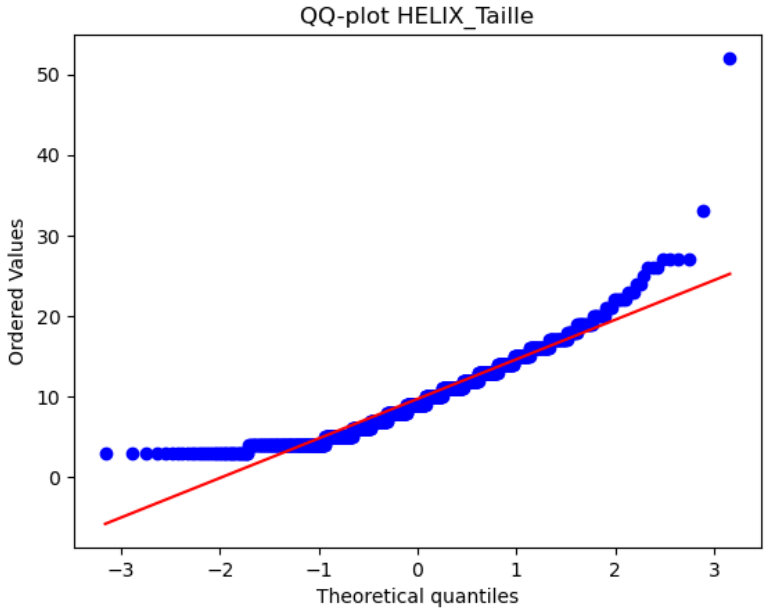
Il y a aussi la ΔCorrélation:

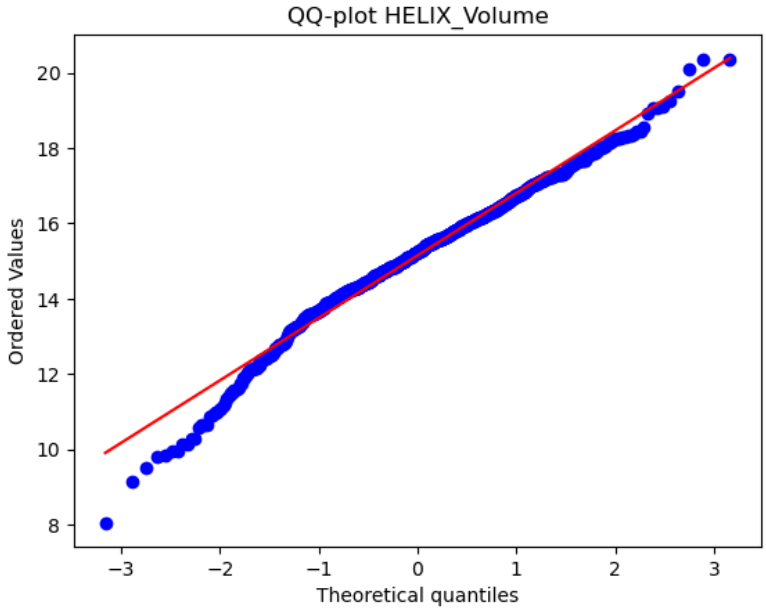


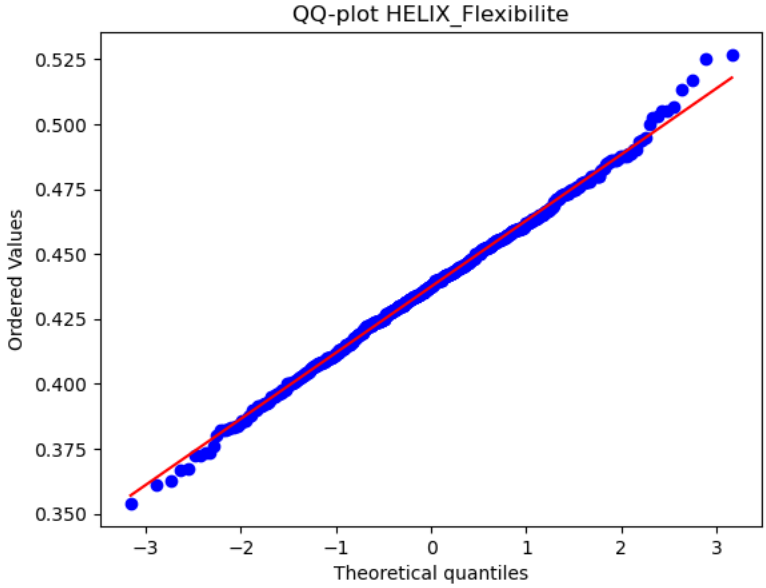
3.3 La normalité

La normalité est vérifiée à l'aide d'un QQ-plot selon une distribution normale:

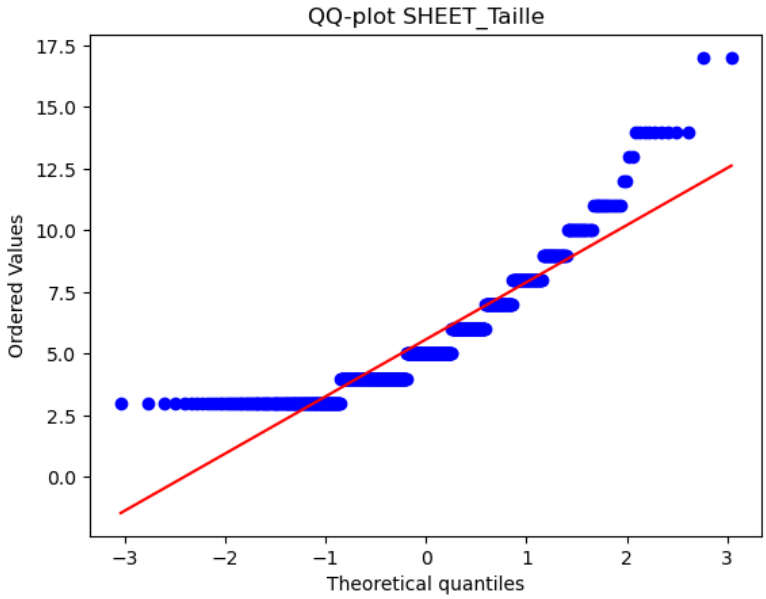
Hélice

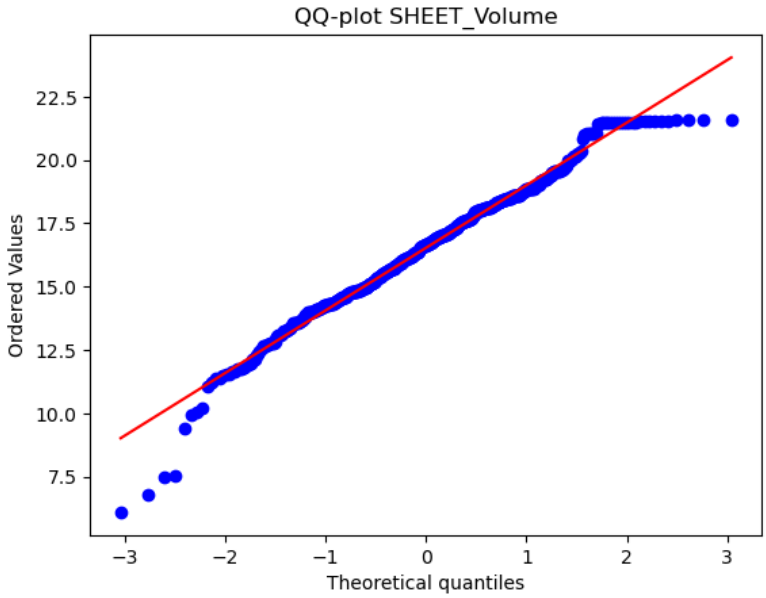


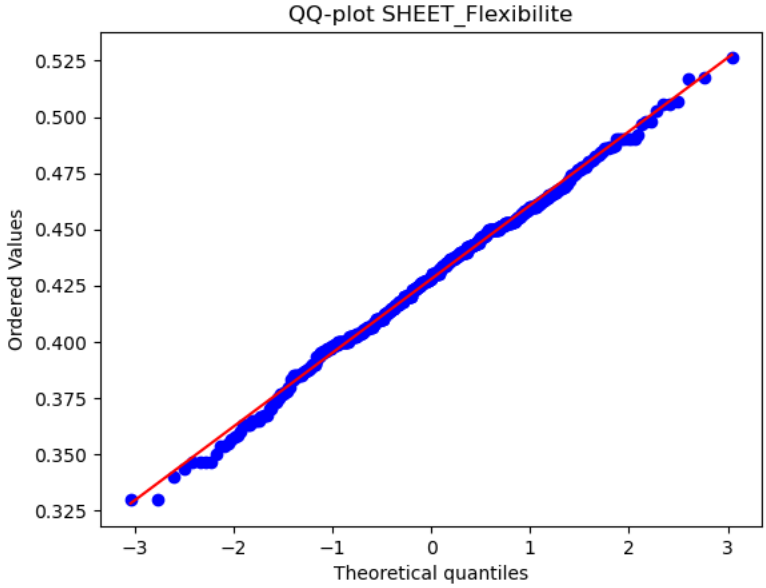



QQ-plot HELIX_Flexibilite
Ordered Values
Theoretical quantiles
0.525
0.500
0.475
0.450
0.425
0.400
0.375
0.350
−3
−2
−1
0
1
2
3

Feuillet


QQ-plot SHEET_Taille
Ordered Values
Theoretical quantiles
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
−3
−2
−1
0
1
2
3


QQ-plot SHEET_Volume
Ordered Values
Theoretical quantiles
22.5
20.0
17.5
15.0
12.5
10.0
7.5
−3
−2
−1
0
1
2
3


QQ-plot SHEET_Flexibilite
Ordered Values
Theoretical quantiles
0.525
0.500
0.475
0.450
0.425
0.400
0.375
0.350
0.325
−3
−2
−1
0
1
2
3

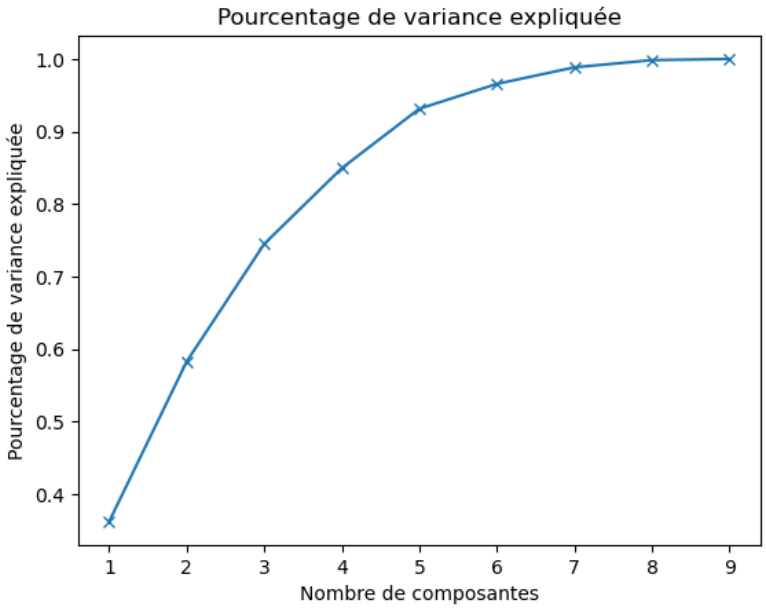
3.4 L'ACP et les variables sélectionnées

Les résultats de l'analyse en composantes principales permettent de sélectionner les variables parmi p variables afin de conserver le maximum d'inertie, défini par :

$$I_N = \frac{1}{2p^2} \sum_{i=1}^{p} \sum_{j=1}^{p} d^2(i,j) \tag{5}$$

On notera l'inertie d'un axe(i): λ_i et sa part d'inertie expliquée $\frac{\lambda_i}{p}$. Avec λ_i la i-ème plus grande valeur propre de la matrice centrée.

On a donc pour le plan axe(i) et axe(j): $\frac{\lambda_i + \lambda_j}{p}$



Out[20]:

	Taille	Polarite	Charge	Volume	Tendance helice	Tendance feuillet	Flexibilite	Hydropathie	Pi
0	-0.170855	-0.380633	0.060687	0.422195	0.039813	0.482591	-0.384782	0.511120	-0.014703
1	-0.071385	-0.055511	-0.695215	-0.054154	0.062277	-0.057648	-0.022910	0.074225	-0.701663
2	-0.264225	-0.358050	0.043674	-0.187087	-0.768920	0.106617	0.395715	0.049841	-0.058279
3	0.816958	-0.441353	0.053343	-0.268889	-0.074806	-0.101772	-0.133393	0.160035	-0.057292
4	0.472414	0.349110	-0.133172	0.513274	-0.188105	0.403812	0.412757	-0.047638	-0.051736
5	0.054930	0.077579	-0.019824	0.310795	-0.506400	-0.183215	-0.633281	-0.444627	-0.072322
6	0.042297	0.488525	-0.087067	-0.574465	-0.198888	0.495727	-0.324383	0.167691	0.057601
7	-0.022253	-0.330728	0.130030	-0.146444	0.257300	0.543278	0.014090	-0.675855	-0.182857
8	0.001057	-0.230458	-0.682367	-0.004080	0.010860	0.067308	-0.014063	-0.142436	0.675361

Enfin, à l'aide d'une méthode brute force, j'ai pu établir les variables permettant de mieux distinguer Hélice et Feuillet. Ces résultats sont en concordance avec les résultats de l'ACP :

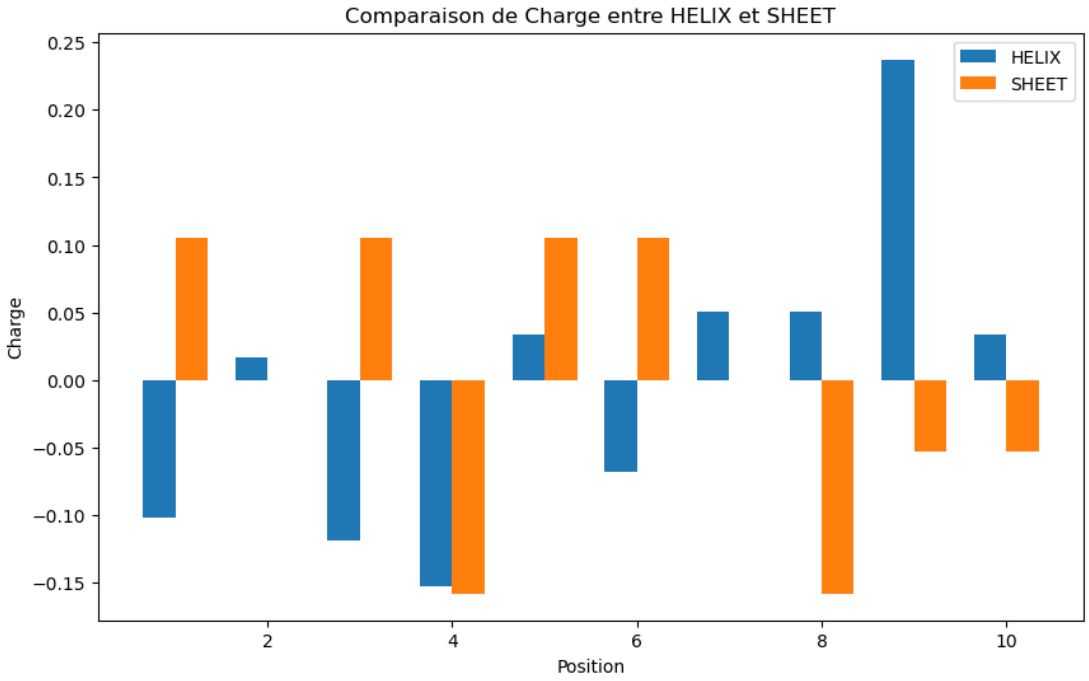
('Taille', 'Charge', 'Tendance helice', 'Tendance feuillet', 'Hydropathie')

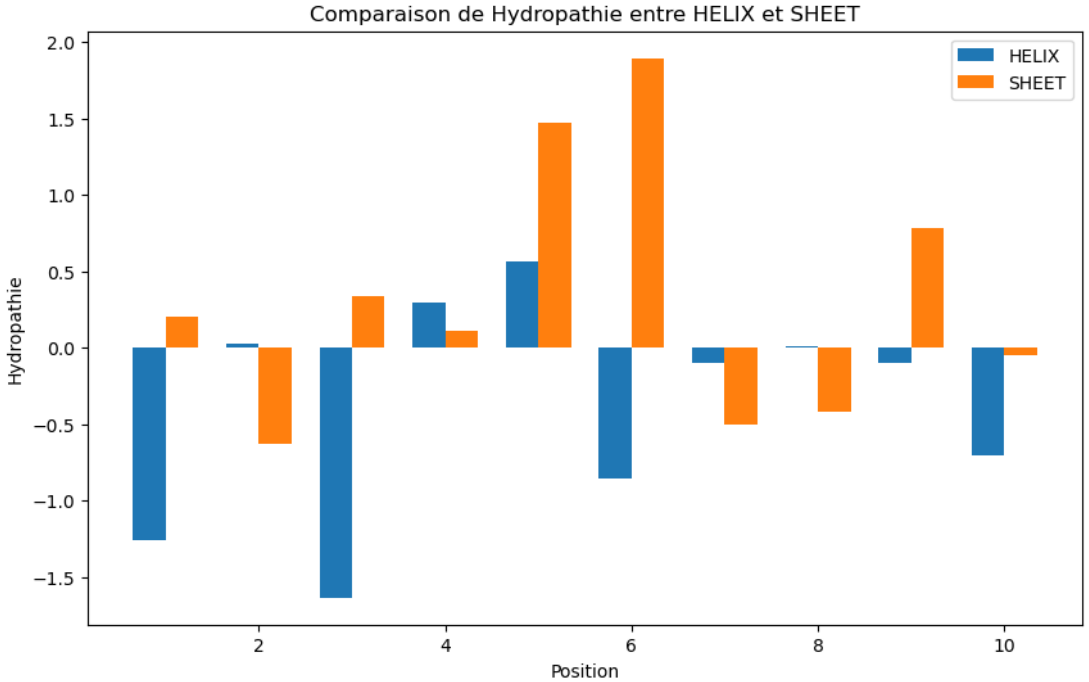
3.5 Comparaison Hélice vs Feuillet

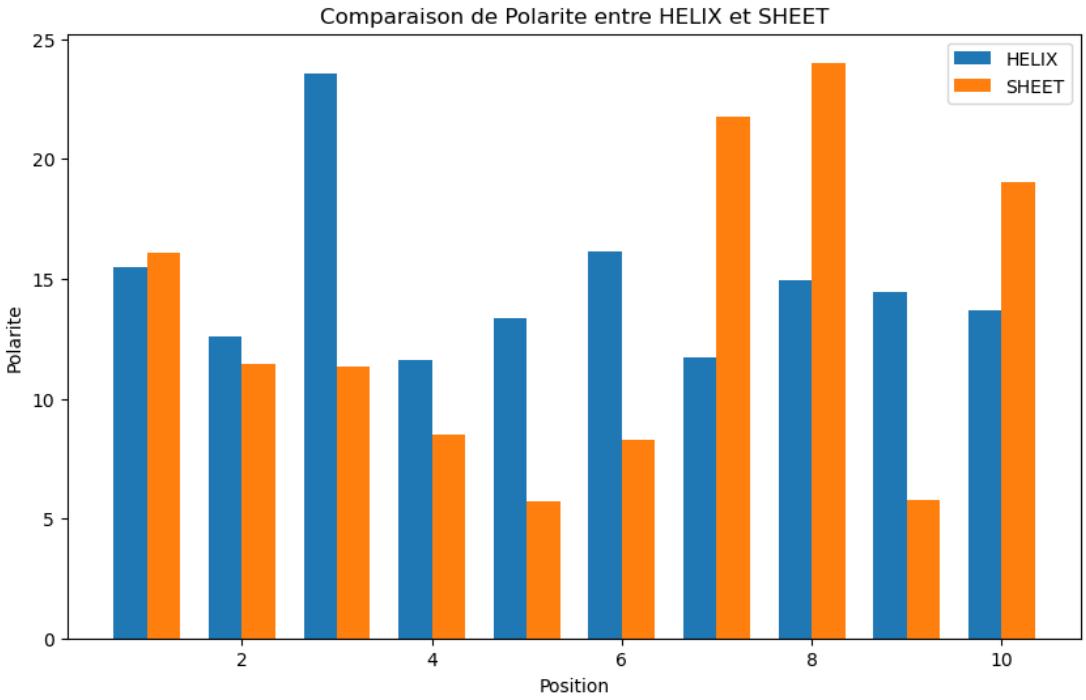
La comparaison des hélices et feuillets de de même taille a été réalisé en faisant la moyenne des variables pour chaque AA à la position i, comme suit:

$$\text{Moy(i, var, str)} = \frac{\sum_{SEQ \in str} SEQ_i[var]}{|str|} \tag{6}$$

Dans le cas suivant pour des Hélices et Feuillets de taille 10:


Comparaison de Charge entre HELIX et SHEET
HELIX
SHEET
Charge
Position
0.25
0.20
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
2
4
6
8
10


Comparaison de Hydropathie entre HELIX et SHEET
HELIX
SHEET
Hydropathie
Position
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
2
4
6
8
10


Comparaison de Polarite entre HELIX et SHEET
HELIX
SHEET
Polarite
Position
0
5
10
15
20
25
2
4
6
8
10

3.6 Clustering et K-means

L'algorithme des K-means a été utilisé, ainsi que leurs qualités pour chaque k. On note µ[d] le centroïde dans la dimension d. Pour chaque itération, on a :

$$\mu[\mathrm{d}] = \frac{\sum_{point \in \mu} point[d]}{|\mu|} \tag{7}$$

On calcule aussi la qualité de chaque clusters comme suit:

$$\text{Homogénéité ou } T_\mu = \frac{\sum_{point \in \mu} d(point, \mu)}{|\mu|} \tag{8}$$

$$\text{Séparabilité ou } S_{\mu 1, \mu 2} = d(\mu 1, \mu 2) \tag{9}$$

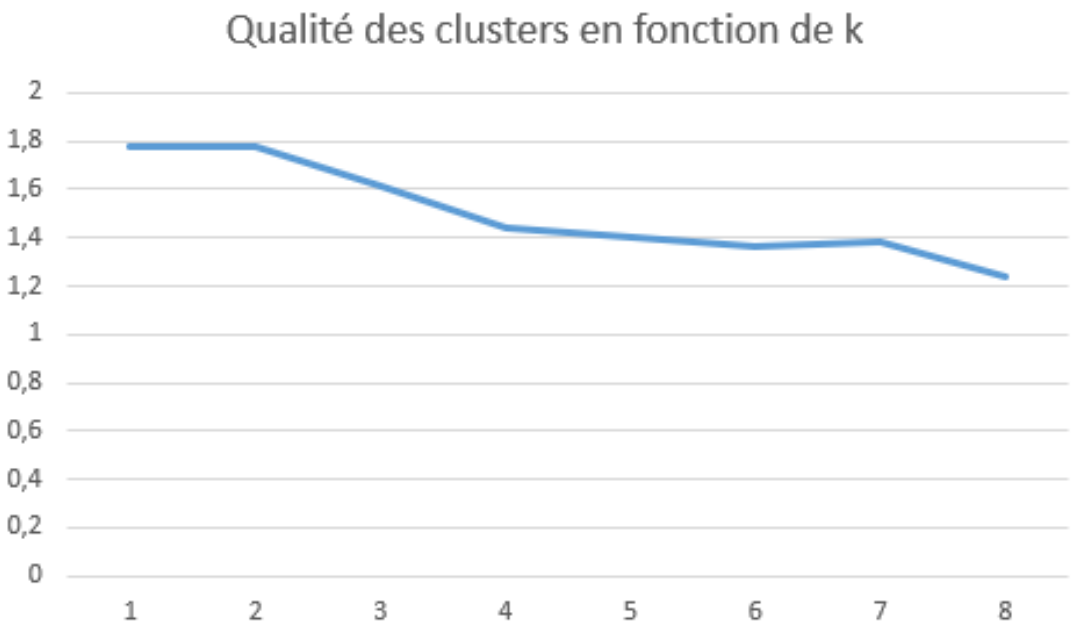
$$\text{Davies Bouldin ou } D_\mu = \max_{\mu 1 \neq \mu 2} \frac{T_{\mu 1} + T_{\mu 2}}{S_{\mu 1, \mu 2}} \tag{10}$$

On a donc la qualité de notre clustering par:

$$\mathrm{D} = \frac{\sum_{i=1}^{K} D_{\mu 1}}{K} \tag{11}$$

On a ainsi pour une partition à K = 2, D = 1.77 et:

	Hélice	Feuillet
$\mu 1$	414	265
$\mu 2$	465	322


Qualité des clusters en fonction de k
2
1,8
1,6
1,4
1,2
1
0,8
0,6
0,4
0,2
0
1
2
3
4
5
6
7
8

3.7 Clustering et CAH

L'algorithme de la CAH a été utilisé avec la dissimilarité Max ainsi on définit la distance entre $\mu 1$ et $\mu 2$ comme suit:

$$\mathrm{d}(\mu 1,\, \mu 2) = \max_{\forall p1 \in \mu 1, \forall p2 \in \mu 2} (d(p1, p2)) \tag{12}$$

Le résultat n'est pas très éloigné d'un 2-means, pour un découpage en 2 groupes (cf: svg).

3.8 Passage en qualitatif

Sur l'ensemble des variables, les acides aminés ont été étiquetés par des variables qualitatives selon le raisonnement suivant :

1. Si AA[var] $\geq$ Quantile(0.8, var): AA[var] = "+ +"
2. Sinon Si AA[var] $\geq$ Quantile(0.6, var): AA[var] = "+"
3. Sinon Si AA[var] $\geq$ Quantile(0.4, var): AA[var] = "0"
4. Sinon Si AA[var] $\geq$ Quantile(0.2, var): AA[var] = "-"
5. Sinon AA[var] = "- -"

Cela nous servira pour l'apprentissage supervisé avec ID3.

3.9 Apprentissage supervisé et K-ID3

On définit un K-ID3 comme un arbre de décision ayant appris sur la séquence avec une fenêtre de [n-k, ..., n, ..., n+k] pour chaque n. Ainsi, pour la séquence MAAKLASS et pour la lysine K, un 2-ID3 a accès aux caractéristiques des AA suivants : [A, A, K, L, A]. Pour tout AA tel que la position est inférieure à n, on prend en compte la prédiction déjà réalisée. Ainsi, le 2-ID3 apprend sur une fenêtre de 2 avec la structure (Vide ou Structure) déjà connue des deux alanines précédentes.

Lors de la prédiction, elle se fait ainsi du Nter vers le Cter, et le K-ID3 prend en compte la décision réalisée (Vide ou Structure) ainsi que les variables des [n-k, n-1] AA et seulement les variables pour les [n, n+k] AA.

Si elles sont dans une structure, le 2-ID3 le sait. Même raisonnement pour 1-ID3, sauf que la fenêtre est [A, K, L] et la structure (Vide ou Structure) de A seulement est connue.

Le but du K-ID3 est donc de prédire les sous-séquences étant des structures.

$$\text{K-ID3}([\mathrm{AA}_{n-k}, \ldots, \mathrm{AA}_n, \ldots, \mathrm{AA}_{n+k}]) \rightarrow \{Vide, Structure\} \tag{13}$$

Ainsi, on obtient les sous-séquences contiguës, qu'on utilisera pour le K-NN.

Lorsque les structures sont données, évitant ainsi de s'auto-tromper, les arbres réalisent un résultat d'environ 80

Cependant, la prédiction ne se fait pas dans ce contexte et doit prendre en compte les anciennes prédictions.

3.10 Classification, K-NN et K-Homologie

L'algorithme des K-NN a été utilisé, et par une technique de brute force, il a été calculé que le nombre de voisins pour les dimensions calculées était 7. Il y avait 82% de bonnes prédictions.

La méthode par K-Homologie consiste à:

1. Calculer la distance entre chaque AA pour les dimensions calculées.
2. Aligner les K-séquences les plus proches en utilisant cette distance pour la substitution, avec une pénalité de gap de -0.5 et le bonus identité de 0.
3. Faire la moyenne des distances entre chaque type de structure.
4. Prendre la structure minimisant la moyenne des distances.

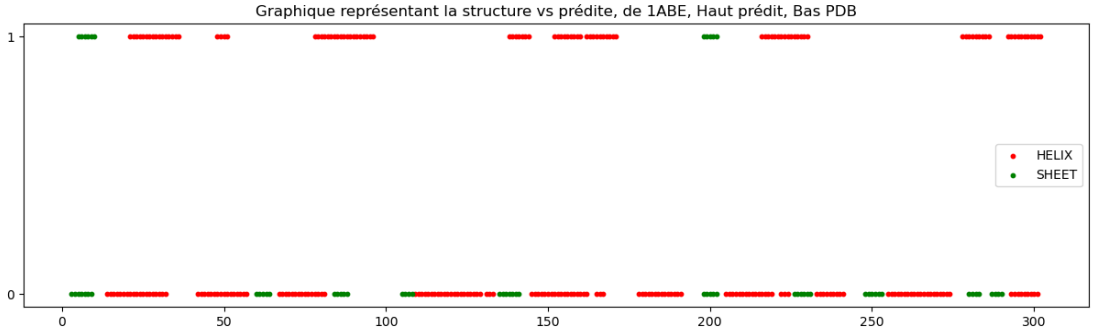
On passe ainsi à l'algorithme, les sous-séquences contigües prédites comme **Structure**.

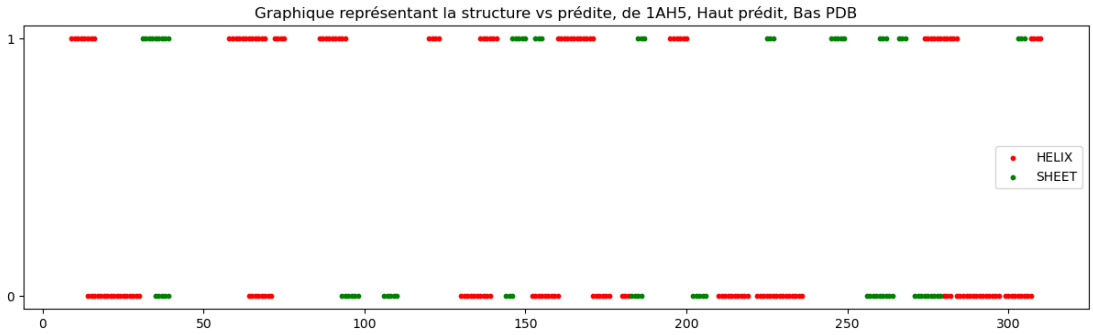
$$\text{K-NN(Sous-Séquence)} \rightarrow \{\text{Hélice}, \text{Feuillet}\} \tag{14}$$

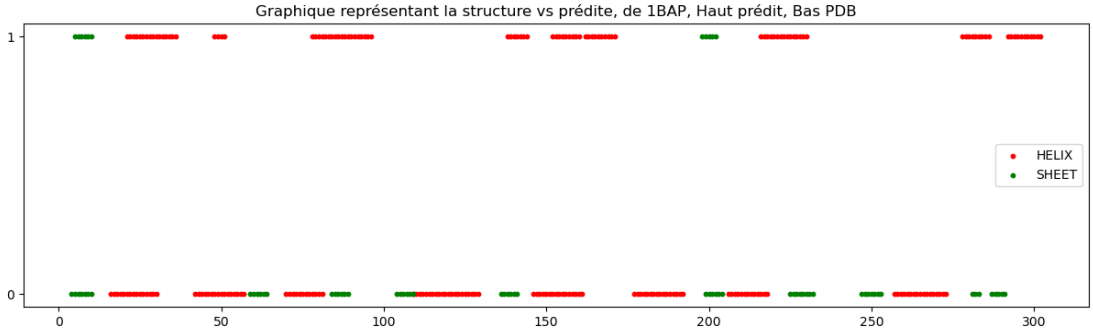
3.11 K-NN(K-ID3(SEQ))

Le K-ID3 prédit ainsi les séquences étant des structures, et le K-NN prédit la structure à partir de la moyenne des variables de la séquence.

Le résultat de la combinaison des deux est le suivant.

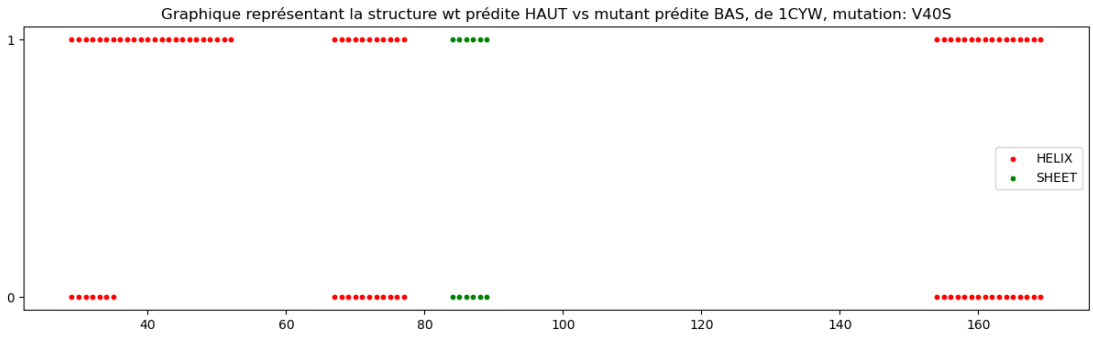



Graphique représentant la structure vs prédite, de 1AH5, Haut prédit, Bas PDB
1
0
0
50
100
150
200
250
300
HELIX
SHEET


Graphique représentant la structure vs prédite, de 1BAP, Haut prédit, Bas PDB
1
0
0
50
100
150
200
250
300
HELIX
SHEET

3.12 Mutation induite

La substitution a été réalisée et son effet peut entraîner des modifications de structures, telles qu'une transition de feuillet à hélice ou encore raccourcir la structure.



3.13 Forêt

L'utilisation d'une forêt ID3 obtient de très bons résultats, mais là aussi en possédant les structures déjà connues de [n-k, n-1] AA, avec un résultat d'environ 90

Cependant, vu que le résultat est la prédiction majoritaire, sur 100 arbres, il est difficile que la prédiction de structure atteigne ≥ 0.5. Ainsi, les résultats sont peu probants.

4 Conclusion et limite

Nous discuterons dans cette partie des limites du projet.

4.1 Les paramètres

Il est très difficile de déterminer les bons paramètres, comme par exemple le bon K pour ID3, le bon score pour l'homologie, la bonne taille de forêt ou encore le bon sous-ensemble de données d'apprentissage pour les K-ID3, car le dataset compte plus de 24 000 individus, et l'entraînement dans ce projet se fait au maximum sur 500 individus parmi plus de 24 000.

Ainsi, cette combinaison de choix rend le bon modèle très difficile à trouver parmi la multitude de choix. Cependant, je reste convaincu qu'il en existe un très performant dans le lot.

4.2 Les données

De plus, les données sont exclusivement des protéines d'*Escherichia coli*, et le jeu de test ne contient pas de protéines d'un taxon différent.

4.3 Amélioration

Pour les pistes d'amélioration, il serait possible de :

1. Coupler le K-Homologie et le K-NN afin de réaliser une prédiction commune.
2. Ajouter des séquences sans structure pour le K-NN, pour rattraper l'erreur d'ID3 si les plus proches voisins n'ont pas de structure.
3. Utiliser un autre algorithme d'apprentissage automatique.

4.4 Conclusion

En somme, les résultats sont très peu probants mais restent cependant intéressants, notamment avec les mutations ou encore les structures prédites. Il est donc possible de mettre en exergue les acides aminés jouant un rôle essentiel dans la structure et, avec une vision très optimiste, prédire l'effet de mutations chez des individus porteurs de maladies génétiques par un simple séquençage.

Bien évidemment, ce projet ne se concentre que sur deux types de structures. Il reste de plus les autres niveaux de conformation bien plus complexes à prédire. Néanmoins, cela ouvre des perspectives pour la suite, en utilisant ces prédictions pour passer à un niveau de conformation suivant, jusqu'à la protéine totalement repliée et dans son milieu.